FEATURES OF STANDARDIZED INSTRUMENTS IN QUALITATIVE EDUCATIONAL ASSESSMENT

Emil LAZĂR[1], Liliana-Maria TODERIUC[2]

Abstract

Among the operations of educational assessment, the decision is based on measurement and appreciation. The discussions about the measurement strategies of educational results, focus on relevant dimensions, how they are being offered by the respondent (separately or globally), and what methods should be used, how situational variables could be controlled so that the measurement should be accurate and confident/ precise.

The hypothesis of this study comes from the interrogation related to the desire of measuring so accurately. A simple, but a fair answer, would be that measurement should ensure data reproducibility, namely the certainty or confidence that what is submissive to the measurement would give the same results whether it would be accomplished by different evaluators, but with the same measuring instrument or at short periods of time.

Therefore, this study is about the features/ properties (values) of qualitative educational assessment and about the methods they provide.

Key words*: Qualitative measurement methods, Validity, Sensitivity, Accuracy, Fidelity.*

1. The necessary interrogation

The necessary interrogation when talking about measurement dimensions and measurement strategies in educational performances or school progress:

Why should we measure so exactly?

A simple but right answer:

A measuring should ensure data reproducibility, namely certainty or trust that what is subject to measurement would give the same results whether it would be accomplished by different evaluators, but with the same measuring instrument or at short periods of time.

Regarding the measurement strategy, we keep in mind:

- „Which are the relevant dimensions?
- How should it be presented by the respondent (e.g. should it evaluate each dimension separately or globally)?

[1] Senior Lecturer, Teachers Training Department, University of Craiova, Romania, email address: lazaremile@gmail.com, corresponding author.

[2] Ph Dc, University of Bucharest, Romania, email address: ltoderiuc@yahoo.com.

- What assessment method should be preferred (standard interview or the evaluating scale)?
- Sampled populations differ in their preferences?
- How can be controlled the situational variables so that the measurement would be accurate and true?" (Froberg, Kane, 1989, pp. 345-354).

1.1. The main qualitative measurement methods

The main qualitative measurement methods are **based on subjective assessments or on observing the behaviour**, they rely on measuring the scaled phenomenon, namely on the identification of its locating/situation in a continuum that theoretically includes all situations within which the phenomenon or subject can be.

By scaling we obtain the scale of assessment or measurement („rating scale") consisting of an imaginary line with a beginning and an end and well-defined anchors. Applying a rating scale to a subject is to ask the respondent to situate himself on this continuum using an "anchor", namely the point where the subject stops in his evaluation.

Another type of measuring **is based on the standardized collection of opinions, of points of view, of subjects' opinions through questionnaires or interviews.** Here is about a global assessment, without a unity or a magnitude of measurement, an evaluation practice that is achieved through structured interviews and questionnaires of assessment.

Within this type of evaluation, „the information is mainly descriptive and the main purpose of standardized assessment instruments is to convert these descriptions into measurable data that can be susceptible of statistical analysis", (Dennis, Ferguson, Tyrer, 1995, p. 145). This is not always possible or desirable. The written or verbal material collected in such evaluative contexts cannot always be nor should be replaced by numerical assessments. They are valid especially for the so called „measurement of change". Measuring the change is necessary for the assessment of the decisions' effectiveness, therefore the focus in the design and development of qualitative-appreciative evaluating methods, should be on instruments with a larger sensitivity and specificity, capable of detecting the smallest changes in educational behaviour/ phenomenon observed/ studied.

2. Construction rating scales

A rating scale **consists of a series/list of individual items**, „each of them covering a well-defined behaviour/ phenomenon, which is evaluated according to the grade of severity" (Bech, Malt, Dencker, Ahlfors, 1993, p. 372). The purpose for using a scale determines its content in items and the various aspects of the structure and procedural/instrumental of its application.

The observed/evaluated is placed along a physical or quasi-physical behavioural dimension, so that certain mathematical properties could be derived. This way, there may be a zero location/position on the scale and equal evaluating

intervals, where a subject/assessed is placed („anchored"). Through cumulative scaling of each item is obtained a general score at the end of the assessment method.

This cumulative aspect of the assessment for each item distinguishes the evaluating scales from the questionnaires.

The items of the rating scales are established/„extracted", systematically from the respondents by the qualitative method „focus-group" and selected, edited and verified by the evaluator's experiences, but also by confrontation with theoretical data validated/offered by scientific research.

The items selection is achieved according to the next principles:

- „comprehensiveness: the used language must be accepted and appropriate;
- to not contain any wording ambiguities, to not be vague nor to have rarely used terms or any specific jargon (slang)/terminology;
- the item should be simple formulated, to not contain more questions at the same time or more answers at the same time;
- the negative formulated items should be avoided as long as possible;
- items formulation should be as short as possible;
- to have discriminative power, meaning to be able to distinguish the subjects that are differently situated on the „beach/expansion" of the scale;
- the items definition should be exhaustive and mutual exclusive (Guilford's criterion);
- the items formulation shouldn't be offensive or trivial" (Streiner, Norman, and Salvador-Carulla, 1995, p. 34).

Scales may include one item, as in global scales, or more. If a scale has more than 30 items, then one speaks of a questionnaire; some authors call them „inventory" or „checklist", especially when assessing only the presence or the absence of the phenomenon.

The items of the questionnaires have strengths and weaknesses, depending on some criteria:

- „They represent a large sample of objectives (+)
- Represent a large sample of contents (+)
- They show the ability to organize, integrate or to synthetize (+)
- They show the level of originality or innovation (-)
- They provide a potential base of diagnosis (+)
- They require a short time of response (+)
- They don't allow the answer interpretation (+)", (Stoica, 2001, pp. 49-66).

Scaling responses is closely related to choosing the methods through which the answers are to be obtained. The choice of method is dictated by what kind of questions are formulated and what is meant to be measured by them (categorical variables or continuous variables).

For categorical variables are built scales in which the respondent is asked to give a „yes-no"/„true-false" answer, or just to tick a response. These are called nominal rating scale.

For continuous variables, there are three kinds of measurements: direct estimation technique, within which the subject is invited to indicate the answer by

marking on a line or ticking a box; the comparative methods, within which the subject chooses from a number of alternatives that have been previously calibrated; the econometric methods, within which the subject is evaluated through anchoring to extreme moods. More frequent in pedagogical field are the methods of direct estimation, such as the analogue scales and the adjectival scale.

Visual analogue scale is a line of a length on which different anchors are placed, the subject being asked to put an „x" or draw a vertical line in the place corresponding to his mood. This type of scale allows the subject to easily communicate his feelings.

The adjectival scale is much more prevalent within the assessment field and they are focused on the adjectival descriptions of continuous or categorical variables. The assessment is achieved using verbal categories or adjectives previously calibrated which are used to quantify the severity of a pedagogical act. They may contain one or more items. Such scale, with gradual arranged alternatives, are also called Likert scale. It is considered that the best reliability is given by the scale with five alternatives, for example: „always", „often", „neither often, nor seldom", „rarely", „never".

The assessment scales management is generally depending on the type of information needed need to be achieved and on the source of information. Thus, the scales that are based on the interview with the subject are called observation scales. The scales that are completed by the subject, without any help are called self-assessment scales.

One might think that the scales completed by an observer or a third person are superior in terms of reliability of given data, unlike the scales completed by the subject himself.

2.1. The quality parameters of the assessment tools

Within the procedural approach of questionnaires, the intention of drafting good statements, in front of which the respondents feel familiar and comfortable with words and can respond effectively, it was used their language, with simple structure, respecting their understanding level, not a strong, specialized language, but a clearly positive or clearly negative one, to produce information.

At the same time, the intention was to be established the presence or absence (implied) of some characteristics, skills, attitudes and of general but integrative behaviours, without any valuable or simple judgment to be issued by the respondent.

Not in the least, the questionnaires should include, (either as investigative form, for some respondents, or as frequency and usefulness to others), elements that involve understanding the work tasks (prescriptive or exploring), the identification of the procedures and instruments of collecting information and of solving (a solution, waiting), the option motivation.

To have a quality standard of a measuring instrument, we must have a standard, an absolute standard or what is called: „gold standard".

Regarding the quality of the questionnaires as investigative tools – validity, reliability, objectivity, application – we must observe not only the significance of data processing for research, but also for the respondents.

Thus, **the VALIDITY**, the fact that „is measured what is destined to be measured", (Ausubel, Robinson, 1981, p. 678) aims to accurately express certain constructs (motivation, interest for a range of knowledge, involvement, networking), focusing on what the respondents are interested in and seeking to forecast the research results (students' questionnaires before and after the assessment with the help of provided instrument – through levels given by learning results). In this respect, „one should consider the characteristics of the sample examined, the difficulty level of the items, their quality", (Stoica, 2001, pp. 49-66).

We refer to two validity types: content validity and construct validity (the extent to which the test measures the level of intelligence, creativity, logical thinking, motivation, involvement, relating and any other construct). Validity also measures the way the test items correspond to the studied contents and to the behaviours signalled by the objectives.

The validity of measuring instruments guarantees that, though we evaluate subjective variables, the measurement results can be comparable from one subject to another. Therefore, the measurement target must be precisely defined and circumscribed, otherwise, the assessment of the validity of an item would be very hard to evaluate. In fact, validity must be consistent with theoretical constructs that were the base construction of a measuring item.

By testing the validity of a scale or interview, we demonstrate the psychometric properties of that instrument to characterize the assessed subjects. As Landy said: „the validation process is not so directed on test integrity but on the inferences, that can be withdrawn about the subjects' skills which produced the score of the test". In other words, a validation of a scale is a process through which we determine the trust degree about test results and deductions about subjects who have different scores at scales they have been applied to.

Another category of quality attributes of measuring instruments is given by the capacity of a scale or an interview to be more **appropriate** to detection (no only to measuring), for the condition or feature it was destined for. This capacity is determined when there is a comparison standard, namely a "gold standard", (Landy, 1986, pp. 1183-1992).

SENSITIVITY is the capacity of a test to correctly identify a type of knowledge (factual, categorical, procedural, metacognitive).

SPECIFICITY is the ability of a test to highlight the degree of knowledge, comprehension, application, analysis, synthesis, critical judgement only for the knowledge that are to be assessed and not for the others.

ACCURACY is the confidence that can be attributed to a result of a test, an accuracy of more than 80%, showing that within this proportion there have been given positive responses to the questionnaire.

FIDELITY is the quality of a test/trial to produce the same results or results with minimal differences, in the case they were successively used and under the same

examining conditions, there being applied other samples with items of the same value. Some researchers tend to use a wide list of synonyms including: ”objectivity”, ”reproducibility”, ”stability”, ”agreement”, “association”, ”sensitivity”, ”accuracy”.

Measurement fidelity brings information not only about measurement error but also about the variability of the measured subjects.

Though it is believed that fidelity is a measure of a test, it must be said that this parameter is closely related to the population intended to be measured. The confidence coefficient has meaning only when it is applied to a certain population and under some measurement conditions.

The fidelity is desired to be raised, starting from the inner reason: „the utility of the proposed assessment. It is complementary used together with other method”, (Gronlund, 1950, pp. 197-225).

Objectivity is the feature that measure the level of agreement among the estimations of more evaluators, concerning the quality of an answer given to each item of a test. The highest objectivity of a test makes it becomes a standardized test.

The quality and applicability of questionnaires is achieved by concordance between their form and content and the respondents' understanding level. The questionnaires quality of applicability is achieved by the concordance between their form and content and the understanding level of the respondents.

Applicability seeks the concordance between the level of knowledge and understanding of those evaluated and the content of the assessment test. It also considers the quality of the sample to be analysed, interpreted and easily assessed, but also the ratio of the items and the evaluation objectives.

Conclusions on choosing an assessment item

Even if the advantages of using standardized instruments are obvious, the selection process of a measuring instrument, isn't always simple. In its selection, there are more general criteria such as:

- The item should produce useful information for evaluation.
- The instrument should produce quantifiable information and if possible, scores that can be used in comparing the subject status over time (individual progress) and other subjects (criterial assessment). Such scores are useful in facilitating a statistical analysis of the performance of an evaluated group.
- The instrument to be easy to administer and not excessively long.
- Language and words used in expressing the items to be appropriate to the cultural and intellectual status of those evaluated, to be acceptable to them and not offensive.
- The instrument to be enough sensitive to the problems raised by those assessed, so that their significant changes could be evaluated.
- The costs of the materials required for the instrument usage not to be too high.

REFERENCES

1. Ausubel, D., Robinson, F. (1981). *Învățarea în Școală*, Bucharest: Didactical and Pedagogical Publisher.
2. Bech, P., Malt, U.F., Dencker, S.J. and U. G. Ahlfors, U.G. (1993). Scales for Assessment of Diagnosis and Severity of Mental Disorders, *Acta Psychiatrica*, Scand., 87, Suppl. p. 372.
3. Dennis, M., Ferguson, B., Tyrer, P. (1995). Rating instruments, in *Research Methods in Psychiatry*, C. Freeman & P. Tyrer (eds.). London: Gaskell.
4. Froberg, D.G., Kane, R.L. (1989). Methodology for measuring health-state preferences - I: Measurement strategies, *J. Clin. Epidemiol., 42*, 345-354.
5. Gronlund, N.E. (1950). The Accuracy of Teachers' Judgments Concerning the Sociometric Status of Sixth-Grade Pupils: Part I, *Sociometry*, American Sociological Association, 13(3) (Aug., 1950), pp. 197-225.
6. Landy, F.J. (1986). Stamp collecting versus sciences, *American Psychologist*, *41*, 1183-1192a.
7. Salvador-Carulla, L. (1996). Assessment instruments in psychiatry: description and psychometric properties, in G. Thornicroft & M. Tansella (Eds.). *Mental Health Outcome Measures*. Berlin: Springer.
8. Stoica, A. (2001). *Evaluarea școlară și examenele,* SNEE. Bucharest: ProGnosis Publisher.
9. Streiner, D.L., Norman, G.R. (1995). *Health Measurement Scales: A Practical Guide to Their Development and Use*. 2nd ed. Oxford: Oxford University Press.